DOCUMENT RESUME

ED 464 918 TM 033 848

AUTHOR Zumbo, Bruno D.; Ochieng, Charles O.

TITLE The Effects of Various Configurations of Likert, Ordered Categorical, or Rating Scale Data on the Ordinal Logistic Regression Pseudo R-Squared Measure of Fit: The Case of the Cummulative Logit Model.

PUB DATE 2002-04-00

NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Educational Research; *Goodness of Fit; *Likert Scales; Monte Carlo Methods; *Rating Scales; *Regression (Statistics); Simulation

IDENTIFIERS *Categorical Data; Logit Analysis

ABSTRACT

Many measures found in educational research are ordered categorical response variables that are empirical realizations of an underlying normally distributed variate. These ordered categorical variables are commonly referred to as Likert or rating scale data. Regression models are commonly fit using these ordered categorical variables as the criterion (i.e., dependent or response) variable; however, a common recommendation in the methodological literature is that researchers make use of ordinal logistic regression when they have these ordered categorical response variables. An advantage of ordinal logistic regression is that it provides a pseudo R-squared measure of fit so that researchers may find this regression model familiar and hence appealing. This study investigated how the pseudo R-squared fit statistic in ordinal logistic regression operates under a variety of conditions over a varying number of Likert scale points and skewness of the Likert data. The study also demonstrates how the regular ordinary least-squares R-squared statistic operates in the same conditions as the Likert data. The pseudo R-squared fit statistic operates well in a majority of conditions, and so it is recommended that educational researchers begin to explore the use of ordinal logistic regression in their modeling practice with ordered categorical data. Using the pseudo R-squared index will ease the transition to ordinal logistic regression because it provides a sense of familiarity to the researcher. (Contains 4 figures, 11 tables, and 6 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 918

The Effects of Various Configurations of Likert, Ordered Categorical, or Rating Scale Data on the Ordinal Logistic Regression Pseudo R-squared Measure of Fit: The case of the cumulative logit model

Bruno D. Zumbo
University of British Columbia

Charles O. Ochieng
CTB/McGraw-Hill

Paper presented in the symposium "Monte Carlo Studies of Statistical Procedures" at the American Educational Research Association conference April 2002, New Orleans.

Address Correspondence to:
Professor Bruno D. Zumbo
University of British Columbia
Scarfe Building, 2125 Main Mall
Department of ECPS
(Program Area: Measurement, Evaluation, & Research Methodology)
Vancouver, B.C.
CANADA V6T 1Z4

Program Area Coordinator: Measurement, Evaluation, & Research Methodology
Associate Member: Department of Statistics

e-mail: bruno.zumbo@ubc.ca

web page: http://www.educ.ubc.ca/faculty/zumbo/

Phone: (604) 822-1931
Fax: (604) 822-3302

TM033848

B. Zumbo

BEST COPY AVAILABLE 2

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Abstract

Many measures found in educational research are ordered categorical response variables that are empirical realizations of an underlying normally distributed variate. These ordered categorical variables are commonly referred to as Likert or rating scale data. Interestingly, regression models are commonly fit using these ordered categorical variables as the criterion (i.e., dependent or response) variable; however, a common recommendation in the methodological literature is that researchers make use of ordinal logistic regression when they have these ordered categorical response variables. An advantage of the ordinal logistic regression is that it provides a (pseudo) R-squared measure of fit so that researchers may find this regression model familiar and hence appealing. The present study investigated how the pseudo R-squared fit statistic in ordinal logistic regression operates under a variety of conditions over varying number of Likert scale points, and skewness of the Likert data. Along the way, we also demonstrated how the regular ordinary least-squares R-squared statistic operates in the same conditions with the Likert data. The pseudo R-squared fit statistic operated well in a majority of the conditions and so it is recommended that educational researchers begin to explore the use of ordinal logistic regression in their modeling practice with ordered categorical data. Using the pseudo R-squared index will ease the transition to ordinal logistic regression because it provides a sense of familiarity to the researcher.

The Effects of Various Configurations of Likert, Ordered Categorical, or Rating Scale Data on the Ordinal Logistic Regression Pseudo R-squared Measure of Fit: The case of the cumulative logit model

Many of the measures obtained in educational research are ordered categorical responses on questionnaires, measures, or surveys. These variates are referred to by different names such as ordered categorical, Likert, or rating scale variables. They are considered ordered-categorical observed variables where the underlying variable is completely unobserved (i.e., latent). Furthermore, as the normally distributed latent variate increases beyond certain threshold values, the observed variable takes on higher scores, referred to as scale points. As is commonly found in the educational research literature, we will refer to these types of variables as Likert variables wherein, for example, such a variable with four possible observed values is commonly referred to as a "four-point Likert scale". As they commonly are in the educational research community, the terms ordered categorical, rating scale and Likert will be used interchangeably in this paper.

The essential features in the above description are that the observed data are ordered categorical outcomes on a variable, Y, based on an underlying continuous variable, η. For example, for a two-point Likert scale a dichotomous variable is observed as Y=2 when η exceeds some threshold value τ, and as Y = 1 otherwise. Furthermore, it is commonly assumed that the underlying continuous variable, η, is Gaussian in form – i.e. normally distributed. Figure 1 depicts an example of how one can characterize Likert responses. Figure 1 visually depicts how two- through to four-point Likert responses may be conceptualized in the case of equal thresholds along the underlying continuous (in this case, a standard normal) variate. The case depicted in Figure 1 is commonly referred to as the "equal threshold" case because the thresholds equally divide the continuum in Figure 1 upon which the latent variate rests; for practical purposes between –3 and 3. For example, for two-point Likert scale the threshold is at zero whereas for a three-point scale the thresholds are at –1.0 and 1.0. Figure 2 depicts an empirical histogram of observed data that arise from a four-point equal threshold situation. (As we will see in the methodology section, these thresholds need not be at equal intervals and doing so may result in skewed empirical ordered categorical data.) Figure 3 depicts a histogram of empirical response data wherein the thresholds are not at equal intervals along the latent continuum – in this case the thresholds are at –2.4, -1.8, -1.2, -0.60, and 0.0 with a standard normal latent variate.

BEST COPY AVAILABLE

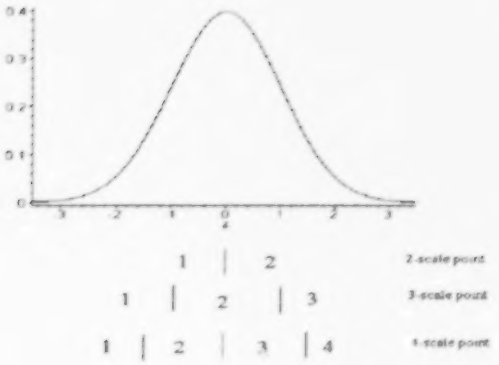

Figure 1. Likert responses from equal interval thresholds.

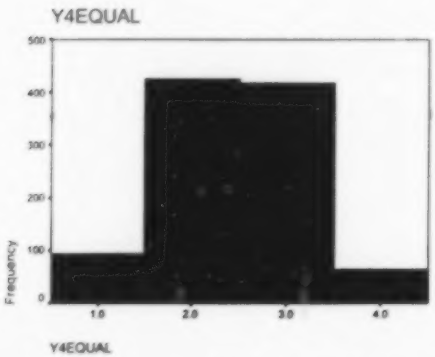

Figure 2. Histogram for 4 scale-point equal interval response pattern.

BEST COPY AVAILABLE

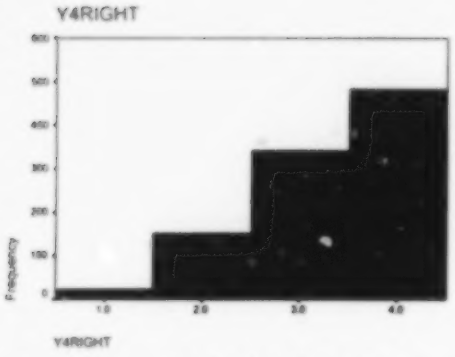

Figure 3. Histogram for 4 scale-point unequal interval response pattern bunched to the right.

Likert variables are not only common in educational and psychological research but they are often used in conventional ordinary least-squares (OLS) regression analyses as dependent variables. Rather than OLS regression using Likert variables (i.e., in essence treating the Likert outcome variables as if they were interval scaled), a common recommendation in the methodological literature is to use statistical methods for ordered categorical data (e.g., Agresti, 1996).

Of the various ordered categorical data methods available the most commonly used one is the so-called "ordinal logistic regression model." This is a multicategory logit model that incorporates the ordering in the dependent variable in the modeling process. The ordinal logistic regression results in both (a) simpler interpretations of the model and results, than the conventional multicategory (i.e., nominal) logistic regression model, and (b) potentially great power than these ordinary multicategory (nominal) logit models (Agresti, 1996, p. 211).

Ordinal Logistic Regression: The cumulative logit model

The interested reader should consult Hosmer and Lemeshow (2000) for a thorough discussion of logistic regression; however, we will provide a brief description of ordinal logistic regression to set the context for our study.

The conventional binary logistic regression procedure uses the response variable (often coded 0 or 1) as the dependent variable, with continuous covariates or categorical factors as explanatory variables. The binary logistic regression equation is commonly written as

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \qquad (1)$$

where Y is a natural log of the odds ratio, α denotes the intercept, and the $\beta_i, i = 1,...,k$, denote the slope coefficients in the multiple regression. Equation (1) could be also written as a parameterization in terms of the logit of Y=1 versus Y=0,

$$\ln\left[\frac{\Pr(Y=1|x_k)}{\Pr(Y=0|x_k)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \qquad (2)$$

where, for example, $\Pr(Y = 1|x_k)$ denotes the probability that Y=1 conditional on the k explanatory variables. This notion of one odds ratio relative to another will be handy in understanding ordinal logistic regression.

The ordinal logistic regression model discussed in this paper can be written as either (a) a regression model that looks much like the ones commonly seen in educational research except it involves an unobserved latent continuum, or (b) as a cumulative logit model. Let us treat each of these in turn.

A. Regression Involving an Unobserved Latent Continuum

One can interpret logistic regression as a linear regression of predictor variables on an unobservable continuously distributed random variable, y^*. Thus, Equation (1) for binary response data can be re-expressed as

$$y^* = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon_i , \qquad (3)$$

where the ε_i are, for the logistic regression model, distributed with mean zero and variance $\pi^2/3$. From this and some additional conditions, one can get a (pseudo) R^2 for ordinal logistic regression (see, Latila, 1993; McKelvey & Zavoina, 1975). It should be noted that the R-squared produced by the ordinal logistic regression in equation (3) is more accurately referred to as a pseudo-R-squared index because it does not share all of the properties of R-squared computed from ordinary least-squares regression. Because the response variable in equation (3) is binary, the intercept, α, does not have a subscript; however, if the response variable is polytomous then there would be more than one intercept term in the model denoted as α_j, j = 1, 2, ..., c-1, where c is the number of categories in the ordinal scale.

It is important to note that the notion of an unobservable continuously distributed random variable is quite familiar in the disciplines of educational and social research. It is simply a latent continuum of variation. We normally conceive of unobserved variables as those which affect observable variables but which are not themselves observable because the observed magnitudes are subject to measurement error or because these variables do not correspond directly to anything that is likely to be measured. Examples are concepts such as "ability", "knowledge", or a personality variable such as "extroversion". Their existence is postulated and they are defined implicitly by the specification of the model and the methods used to estimate it.

B. Cummulative Logit Model

Equation (3) reminds us that ordinal logistic regression assumes an unobservable continuum of variation (i.e., a latent variable). However, ordinal logistic regression can also be expressed as a ratio of logits, much like the binary logistic regression case. That is, rather than reference Y=1 to Y=0, as in equation (2), the ordinal logistic regression can also be expressed as

BEST COPY AVAILABLE

$$\log\left[\frac{\Pr(Y \le j)}{\Pr(Y > j)}\right] = \alpha_j + +\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \text{, or}$$

$$\text{logit}[\Pr(Y \le j)] = \alpha_j + +\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \text{,} \qquad (4)$$

where a logit is the natural logarithm of the ratio of two probabilities as seen in Equation (2), and as seen in equation (3) j = 1, 2, ..., c-1, where c is the number of categories in the ordinal scale. In the language of equation (4), the model requires a separate intercept parameter α_j for each cumulative probability. Equation (4) highlights two assumptions of ordinal logistic regression (Agresti, 1996):

1. It operates on the principle of cumulative information along the explanatory variable (or hyperplane in the case of multiple regression). That is, for example, for a 3-point response an ordinal logistic regression model describes two relationships: the effect of X (in our case the total score for the scale) on the odds that Y ≤ 1 instead of Y > 1 on the scale, and the effect of X on the odds that Y ≤ 2 instead of Y > 2. Of course, for our three point scale, all of the responses will be less than or equal to three (the largest scale point) so it is not informative and hence left out of the model. The model requires two logistic curves (see Figure 4), one for each cumulative logit. The dashed line in Figure 4 depicts the cumulative probability for scoring less than or equal to 1 versus greater than 1; and the solid line depicts the cumulative probability for scoring less than or equal to 2 versus greater than 2.

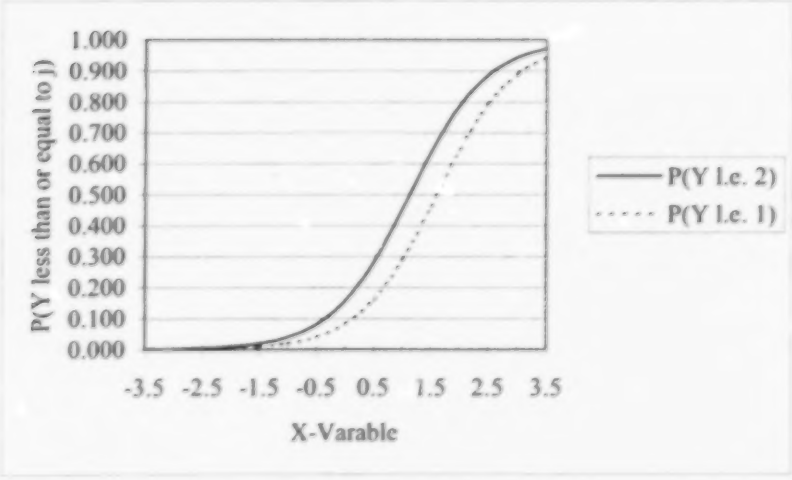

Figure 4. Curves for the cumulative probabilities in a cumulative logit model.

2. At any given point on the X-axis the order of the two logistic curves is the same. For example, in Figure 4 at any point on the X-axis, if one starts at the X-axis going directly upward one will first come across the dashed line, where in the legend to the Figure P(Y le. 1) denotes "the probability that Y is less than or equal to one",

BEST COPY AVAILABLE

and then the solid line, P(Y I.e. 2). This means that at any point on the X-axis the order of the lines are the same. This means that the logistic curves have a common slope, denoted *b* in Equation (4). When interpreting the magnitude of *b* please see Agresti (1996, pp. 213-214).

In summary, if we had for example a 3-point Likert response variable and only one explanatory variable, an ordinal logistic regression models the odds that someone will select the scale point 2 (or less) on the scale in comparison to selecting a response higher on the scale. Furthermore, the regression model does this for each point on the scale simultaneously. What a researchers ends up with is a regression equation having more than one intercept coefficient and only one slope. The common slope assumption could be tested with a nominal multinomial logit model.

As we can see in equation (3) the cumulative logit model with the common slope (i.e., what we will refer to as the ordinal logistic regression model) looks very much like the regular OLS model found in much of regression practice in educational research. In particular, the R-squared index for model fit is intended to give the analyst a sense of the variation accounted for in the latent variable by the k explanatory variables.

The ordinal logistic regression described in this paper is often recommended because it has many common features with standard OLS regression that will help ease educational researchers into adopting this methodology. For example, the presence of an R-squared index may ease educational researchers' shift to ordinal logistic regression modeling.

What is unknown, however, is how does ordinal logistic regression model fit (and particularly the R-squared index) perform under various conditions found with Likert-type data. That is, if the R-squared index in equation (3) is meant to indicate to the researcher how much of the variation in the latent variable is accounted for by the explanatory variables, how does this R-squared index function for (a) a variety of scale points ranging from two to nine, and (b) symmetry versus skew in the observed Likert variable. Ochieng (2001) and others have shown that OLS methods underestimate the R-squared with Likert data. In part, the motivation for this paper is to see how ordinal logistic compares in performance relative to OLS regression on the empirical Likert observations (i.e., treating the Likert variables as interval scale outcomes).

Methodology

This study examined the effect of Likert-type measurement in the criterion (i.e., dependent) variable on the R-squared index of ordinal regression models. The focus of the simulation was: what happens to the model R-squared produced by ordinal logistic regression as we manipulated the (a) number of scale points and (b) response distribution of the Likert dependent variable in the regression.

That is, it is commonly observed in real data that the responses on a Likert-type question are skewed to one end of the scale. Therefore, two types of response patterns with:

1. Equal intervals on the latent continuum, η, resulting in the responses being a symmetric bell-shaped distribution in the center of the response scale, and
2. Unequal intervals on η resulting in positively skewed responses.

BEST COPY AVAILABLE

We also explored negatively skewed responses by using symmetric thresholds on the other side of zero on η but, as expected, the results were precisely the same as the positively skewed responses.

An important methodological point in the simulation is that we wanted to address population analogues of the R-squared measures rather than deal with sample-to-sample variability. The research question we were asking was of the population analogues. To side-step the matter of sample-to-sample variability and focus on large-sample effects, a sample of 50,000 continuous normally distributed scores was generated based on each of three correlation matrices – each with their own different R-squared values that would be obtained from an OLS regression based on that correlation matrix. These normally distributed scores represented the (typically unobserved) latent scores from which the Likert-type responding were simulated.

The simulation was conducted for the case of three predictors and one response (also called criterion or dependent) variable. Only the criterion was be a Likert-type response; in essence, in an educational research context it could be a question on an educational survey. There were eight levels of Likert-type scale points, ranging from two to nine scale points.

The effect of number of scale points and response distribution were examined for three patterns of correlation matrices with low, moderate and high inter-correlation among the variables. The correlation matrices were considered examples of the relationship between predictors and three criterion variables often found in educational settings and social science research (Stevens, 1986). Furthermore, we compared the R-squared produced by ordinal logistic regression to that of the R-squared produced by the OLS regression of the original variables (i.e., the underlying continuous variable).

Given the description above, the study design was a 2x3x8 design with: (a) two types of empirical Likert response distributions, symmetric and skewed, (b) three correlation matrices, and (c) eight Likert scale points ranging from 2 to 9. For each of these 48 finite populations, the continuous scores (which represent the unobserved latent variable) were manipulated and so as to mimic Likert-type responding.

The dependent variables in each cell of the simulation design were (a) the pseudo R-squared produced by the ordinal logistic regression[1] and (b) the R-squared from the regular OLS regression of the Likert dependent variable on the predictors and hence having treated the dependent variable as if it were interval scale. Note that the two regressions described above were computed from the same data; it was just the type of regression that differed.

In essence, the simulation methodology mimicked the process of responding to a Likert-type question and then used the responses as the dependent variable in an ordinal logistic regression, and then again in an ordinary least-squares regression – having treated the dependent variable as interval. As such, the objectives of the simulation were:

(a) To compare the pseudo R-squared produced by ordinal logistic regression to the R-squared that one would have obtained, with the same data, had they been able

[1] Ordinal logistic regression was computed in SPSS with a special macro. Our analysis makes use of a public domain SPSS macro called ologit2.inc (written by Prof. Dr. Steffen Kühnel, and modified by John Hendrickx University of Nijmegen The Netherlands). Write to the authors for a copy or the website where it can be found.

BEST COPY AVAILABLE

to conduct the OLS regression using the continuous latent variable, rather than the Likert-type responses. The expectation from statistical theory is that the pseudo R-squared from the ordinal logistic regression will be close to that of the R-squared using the "latent" continuous variables. This is the expectation because this is precisely the objective of the pseudo R-squared.

(b) To compare the pseudo-R-squared from the ordinal logistic regression with the OLS R-squared that treats the Likert dependent variable as if it were interval scaled. This comparison is made in light of the observation in point (a) above.

At this point we will describe the correlation matrices used in the simulation and the symmetry conditions for the observed variables in more detail.

Correlation Matrices

Correlation matrices selected for simulating the population of responses were based on the typical occurrence in social science research of the relationship among predictors and criterion variables, in which (1) predictors were moderately correlated with each other and low correlation with the criterion, *matrix 1* (2) predictors had a low correlation with each other but a moderate correlation with the criterion, *matrix 2* and (3) predictors had a moderate to high correlation with themselves and moderately high correlation with the criterion variable, *matrix 3*. The correlation matrices were based on examples of the relationship between predictors and criterion variables often found in educational settings and social science research (Stevens, 1986). Table 1 lists the three correlation matrices.

Table 1. Correlation matrices used in the study.

	Y	X_1	X_2	X_3
Y	1			
X_1	.20	1		
X_2	.10	.50	1	
X_3	.30	.40	.60	1

R-squared for the regression of all continuous variables R-squared = 0.118

	Y	X_1	X_2	X_3
Y	1			
X_1	.60	1		
X_2	.50	.20	1	
X_3	.70	.30	.20	1

R-squared for the regression of all continuous variables R-squared = 0.753

BEST COPY AVAILABLE

	Y	X_1	X_2	X_3
Y	1			
X_1	.60	1		
X_2	.70	.70	1	
X_3	.70	.60	.80	1

R-squared for the regression of all continuous variables R-squared = 0.562

Distribution of the Likert variables – symmetry and skew

The symmetric observed Likert variable partitioned the continuous latent variable at equal intervals between –3 and 3 (i.e., note that the latent variable was a standard normal variate). This is the same methodology as that used by Bollen and Barb (1981) in their study of the correlation coefficient. Figure 1 depicts the symmetric case and Table 2 lists the thresholds for the symmetric case.

The asymmetric case followed the same methodology except that the thresholds were selected to be unequal along the latent continuum and hence resulting in a skewed response distribution on the Likert scale. Table 3 lists the threshold for the skewed Likert responses. Tables 4 and 5 list the empirical index of skew for the Likert data, centered at zero.

It is important to note that what was being divided in the response process was not the area under the Normal curve but rather the spatial distance along the continuum. What this represented was the item response model in which the response one provided depended on how much of the latent variable one possessed. For example, having started from the far left and using the 2-point scale, if one only had –1.5 standard units of the latent variable then they responded "1" to the question. On the other hand, in the same context, if one had 0.5 standard units of the latent variable they would have responded "2".

For example, for a two-point Likert scale, the threshold was set at the point, $z = -1.5$, below which the response value was 1 and above which the response value was 2. Similarly, for a three-point Likert scale, two threshold points were created. The two scale points were $z = -1.5$ and $z = 0$. Thus, below $z = -1.5$ on η, the response value was 1, between $z = 0$ and $z = -1.5$, the response values was 2, and above threshold $z = 0$, the response value was 3. In the case of positively skewed responses the threshold points were symmetric above zero as those in Table 2. Because the results were the same for the two types of skew (negative and positive) we did not distinguish between them in reporting the results.

BEST COPY AVAILABLE

Table 4. Skew statistics for the symmetric Likert responses.

	Skewness	Std. Error of Skewness
Y2EQUAL	.005	.011
Y3EQUAL	-.002	.011
Y4EQUAL	-.008	.011
Y5EQUAL	.007	.011
Y6EQUAL	.006	.011
Y7EQUAL	.008	.011
Y8EQUAL	.009	.011
Y9EQUAL	.011	.011

Note: Computed with SPSS

Table 5. Skew statistics for the skewed Likert responses.

	Skewness	Std. Error of Skewness
Y2RIGHT	-3.464	.011
Y3RIGHT	-.599	.011
Y4RIGHT	-.887	.011
Y5RIGHT	-1.045	.011
Y6RIGHT	-1.146	.011
Y7RIGHT	-1.211	.011
Y8RIGHT	-1.261	.011
Y9RIGHT	-1.295	.011

Note: Computed with SPSS

Results and Conclusions

We will present the results for each correlation matrix separately.

***Matrix 1*, Predictors are moderately correlated with each other and low correlation with the criterion**

For this correlation matrix the R-squared for the continuous dependent variable was 0.118. The following observations can be made from Tables 6 and 7:

(a) The pseudo R-squared produced by ordinal logistic regression was, but with one exception, less than the R-squared it was approximating (note that this is not an estimate in the traditional statistical sense but rather an approximation). The odd case was the two-point Likert scale wherein the pseudo R-squared was inflated for the skewed distribution.

(b) Comparison of pseudo R-squared from the ordinal logistic regression with the OLS R-squared coming from treating the Likert variable as if it were interval scaled showed that the difference is quite small in magnitude except for the two-point Likert scale with skew where it was markedly greater than the pseudo R-

squared from the continuous value of 0.118. Likewise the OLS R-squared increased toward the continuous latent variate R-squared (i.e., 0.118) as the number of scale points increased but this change reduced at about 4 points on the Likert scale.

Table 6. Matrix 1 with symmetric Likert responses.

Likert scale points	Pseudo R^2	OLS R^2
2	0.094	0.074
3	0.106	0.084
4	0.105	0.096
5	0.106	0.103
6	0.102	0.106
7	0.105	0.110
8	0.103	0.110
9	0.103	0.112

Table 7. Matrix 1 with negatively skewed Likert responses.

Likert scale points	Pseudo R^2	OLS R^2
2	0.137	0.031
3	0.101	0.085
4	0.099	0.090
5	0.100	0.090
6	0.100	0.091
7	0.100	0.091
8	0.100	0.091
9	0.100	0.091

***Matrix 2*, Predictors have a low correlation with each other but a moderate correlation with the criterion.**

For this correlation matrix the R-squared for the continuous dependent variable wa 0.753. The following observations can be made from Tables 8 and 9:

(a) The pseudo R-squared produced by ordinal logistic regression was, but with one exception, less than the R-squared it was approximating. Again, the odd case was the two-point Likert scale wherein the pseudo R-squared was inflated for the skewed distribution. Overall, however, the pseudo R-squared was a good approximation of the continuous latent variable case it was approximating.

(b) Comparison of pseudo R-squared from the ordinal logistic regression with the OLS R-squared coming from treating the Likert variable as if it were interval scaled showed that the difference is large however this difference decreased with an increasing number of scale points for the Likert variable. As in the previous correlation matrix for the symmetric Likert variate the OLS R-squared increased toward the continuous latent variate R-squared as the number of scale

points increased but this change declined at about 4 points on the Likert scale. For the skewed Likert variable, however, the OLS R-squared treating the Likert variate as if it were interval scaled resulted in a under approximation of the continuous R-squared for the latent variates.

Table 8. Matrix 2 with symmetric Likert responses.

Likert scale points	Pseudo R^2	OLS R^2
2	0.738	0.478
3	0.760	0.563
4	0.745	0.626
5	0.748	0.670
6	0.749	0.695
7	0.748	0.709
8	0.741	0.716
9	0.746	0.726

Table 9. Matrix 2 with skewed Likert responses.

Likert scale points	Pseudo R^2	OLS R^2
2	0.755	0.203
3	0.742	0.553
4	0.743	0.585
5	0.741	0.589
6	0.739	0.590
7	0.742	0.592
8	0.740	0.590
9	0.738	0.585

***Matrix 3,* Predictors have a moderate to high correlation with themselves and moderately high correlation with the criterion variable.**

For this correlation matrix the R-squared for the continuous dependent variable was 0.562. The results are shown in Tables 10 and 11. The conclusions in this case are much the same as in Matrix 2.

Table 10. Matrix 3 with symmetric Likert responses.

Likert scale points	Pseudo R^2	OLS R^2
2	.534	.356
3	.557	.417
4	.553	.466
5	.550	.502
6	.551	.517
7	.545	.528
8	.544	.532
9	.547	.541

Table 11. Matrix 3 with skewed Likert responses.

Likert scale points	Pseudo R^2	OLS R^2
2	.585	.151
3	.545	.412
4	.544	.435
5	.540	.438
6	.541	.439
7	.540	.438
8	.539	.437
9	.540	.435

Recommendations for Educational Data Analysis

The results are somewhat preliminary but we can speak to the research questions we cast at the end of the introduction of this paper.

1. The pseudo R-squared produced in ordinal logistic regression is, in general, a good approximation to the R-squared for the explanatory variables and the latent variate.
2. The pseudo R-squared from ordinal logistic regression does a better job than the OLS R-squared of depicting the R-squared found among the latent variables. That is, the pseudo R-squared from ordinal logistic regression is, in general, closer magnitude to the R-squared found with the latent continuous variables.

Although further methodological research is needed, in the end we are recommending that educational researchers begin to explore the use of ordinal logistic regression in their modeling practices.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.

Bollen, K.A., & Barb, K. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*, 232-239.

Latila, T. (1993). A pseudo-R^2 measure for limited and qualitative dependent variables. *Journal of Econometrics, 56*, 341-356.

McKelvey, R. D., & Zavoina, L (1975). A statistical model for the analysis of ordinal dependent variables. *Journal of Mathematical Sociology, 4*, 103-120.

Ochieng, C. O. (2001). *Implications of Using Likert Data in Multiple Regression Analysis*. Unpublished Doctoral Dissertation, University of British Columbia.

Stevens, J. (1986). *Applied Multivariate Statistics for the Social Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Table 2. Threshold values for the scale intervals in symmetric response distribution with equal interval

1	2	3	4	5	6	7	8	9
0.0000								
-1.0000	1.0000							
-1.5000	0.0000	1.5000						
-1.8000	-0.6000	0.6000	1.8000					
-2.0000	-1.0000	0.0000	1.0000	2.0000				
-2.1429	-1.2857	-0.4286	0.4286	1.2857	2.1429			
-2.2500	-1.5000	-0.7500	0.0000	0.7500	1.5000	2.2500		
-2.3333	-1.6667	-1.0000	-0.3333	0.3333	1.0000	1.6667	2.3333	

Table 3. Threshold values for the scale intervals in negatively skewed response distribution (right bunching) with unequal interval

1	2	3	4	5	6	7	8	9
-1.5000								
-1.5000	0.0000							
-2.0000	-1.0000	0.0000						
-2.2500	-1.5000	-0.7500	0.0000					
-2.4000	-1.8000	-1.2000	-0.6000	0.0000				
-2.5000	-2.0000	-1.5000	-1.0000	-0.5000	0.0000			
-2.5714	-2.1429	-1.7143	-1.2857	-0.8571	-0.4286	0.0000		
-2.6250	-2.2500	-1.8750	-1.5000	-1.1250	-0.7500	-0.3750	0.0000	